Dublin City University at CLEF 2005: Cross-Language Speech Retrieval (CL-SR) Experiments

Adenike M. Lam-Adesina Gareth J. F. Jones
School of Computing, Dublin City University, Dublin 9, Ireland
{adenike,gjones}@computing.dcu.ie

Abstract. The Dublin City University participation in the CLEF 2005 CL-SR task concentrated on exploring the application of our existing information retrieval methods based on the Okapi model to the conversational speech data set. This required an approach to determining approximate sentence boundaries within the free-flowing automatic transcription provided to enable us to use out summary-based pseudo relevance feedback (PRF). We also performed exploratory experiments on the use of the metadata provided with the document transcriptions. Topics were translated into English using Systran V3.0 machine translation. In most cases Title field only topic statements performed better than combined Title and Description topics. PRF using our adapted methods is shown to be affective, and absolute performance is improved by combining the automatic document transcriptions with additional metadata fields.

1 Introduction

The Dublin City University participation in the CLEF 2005 CL-SR task [1] concentrated on exploring the application of our existing information retrieval methods based on the Okapi model to this data set, and exploratory experiments on the use of the provided document metadata. Our official submissions included both the English monolingual and French bilingual runs. This paper reports additional results for German and Spanish bilingual runs. Topics were translated into English using the Systran V3.0 machine translation system. The resulting English topics were applied to the English document collection.

Our standard Okapi retrieval system incorporates a summary-based pseudo relevance feedback (PRF) stage. This PRF system operates by selecting topic expansion terms from document summaries, full details are described in [2]. However, since the automated transcriptions of the conversational speech documents do not contain punctuation, we needed to develop a method of selecting significant document segments to identify documents "summaries". Details of our method for doing this are described in Section 2.1.

The spoken document transcriptions are provided with a rich set of metadata, further details are available in [1]. It is not immediately clear how best to exploit this most effectively in retrieval. This paper reports our initial exploratory experiments in making use of this additional information by merging it with the standard document transcriptions.

The remainder of this paper is structured as follows: Section 2 overviews our retrieval system and describes our sentence boundary creation technique, Section 3 presents the results of our experimental investigations, and Section 4 concludes the paper with a discussion of our results.

2 System Setup

The basis of our experimental system is the City University research distribution version of the Okapi system [3]. The documents and search topics are processed to remove stopwords from a standard list of about 260 words, suffix stripped using the Okapi implementation of Porter stemming [4] and terms are indexed using a small standard set of synonyms. None of these procedures were adapted for the CLEF 2005 CL-SR test collection.

2.1 Term Weighting

Document terms were weighted using the Okapi BM25 weighting scheme developed in [3] calculated as follows,

$$cw(i,j) = \frac{cfw(i) \times tf(i,j) \times (K1+1)}{K1*((1-b)+(b \times ndl(j))) + tf(i,j)}.$$

where *cw(i,j)* represents the weight of term *i* in document *j*, *cfw(i)* is the standard collection frequency weight, *tf(i,j)* is the document term frequency, and *ndl(j)* is the normalized document length. *ndl(j)* is calculated as *ndl(j) = dl(j)/avdl* where *dl(j)* is the length of *j and avdl* is the average document length for all documents. *k1* and *b* are empirically selected tuning constants for a particular collection. *k1* is designed to modify the degree of effect of *tf(i,j)*, while constant *b* modifies the effect of document length. High values of *b* imply that documents are long because they are verbose, while low values imply that they are long because they are multi-topic. The values used for out submitted runs were tuned using the provided training topics.

2.2 Pseudo-Relevance Feedback

We apply PRF for query expansion using a summary-based method described in [2] which has been shown to be effective in our previous submissions to CLEF, including [5] and elsewhere. The main challenge for query expansion is the selection of appropriate terms from the assumed relevant documents. Our query expansion method selects terms from summaries of the top ranked relevant document. All non-stopwords in the summaries are ranked using a slightly modified version of the Robertson selection value (rsv) [3] shown in equation (1).

$$rsv(i) = r(i) \times rw(i). \quad (1)$$

where *r(i)* = the total number of relevant documents containing term *i*, and *rw(i)* is the standard Robertson/Sparck Jones relevance weight [3],

$$rw(i) = \log \frac{(r(i)+0.5)(N-n(i)-R+r(i)+0.5)}{(n(i)-r(i)+0.5)(R-r(i)+0.5)}$$

where *r(i)* = is defined as before, *n(i)* = the total number of documents containing term *i*, *R* = the total number of relevant documents for this query, and *N* = the total number of documents

The top ranked terms are then added to the topic. In our modified version of *rsv(i)*, potential expansion terms are selected from the summaries of the top ranked documents, but ranked using statistics from a larger number of assumed relevant ranked documents from the initial run.

2.2.1 Sentence Selection

Our standard process for summary generation is to select representative sentences from the document [6]. Since the transcriptions in the CL-SR document set do not contain punctuation marking, we needed an alternative approach to identifying significant units in the transcription. We approached this using a method derived from Luhn's word cluster hypothesis. Luhn's hypothesis states that significant words separated by not more than 5 non-significant words are likely to be strongly related. Clusters of these strongly related word were identified in the running document transcription by searching for word groups separated by not more than 5 insignificant words, as shown in Figure 1. Note that words appearing between clusters are not included in clusters, but can be ignored for the purposes of query expansion since they are by definition stop words.

> … this chapter gives a brief description of the **[***data* sets used in *evaluating* the *automatic* relevance *feedback* procedure *investigated* in this *thesis*] and also discusses the extension of …

Fig 1. Example of Sentence creation

The clusters were then awarded a significance score based on two measures.

Luhn's Keyword Cluster Method Luhn's method assigns a sentence score for the highest scoring cluster within a sentence. We adapted this method to assign a cluster score as follows:

$$SS1 = \frac{SW^2}{TW}.$$

where *SS1* = the sentence score
SW = the number of bracketed significant words
TW = the total number of bracketed words

For the examine in Fig. 1, *SW*=6 and *TW*=14.

Query-Bias Method This method assigns a score to each sentence based on the number of query terms in the sentence as follows:

$$SS2 = \frac{TQ^2}{NQ} \ .$$

where *SS2* = the sentence score
TQ = the number of query terms present in the sentence
NQ = the number of terms in a query

The overall score for each sentence (cluster) was then formed by summing these two measures for each sentence.

3 Experimental Investigation

This section describes the establishment of the parameters for our experimental system and then gives results from our investigations.

3.1 Selection of System Parameters

In order to set the appropriate parameters for our feedback runs, we carried out development runs using the CLEF 2005 CL-SR training topics. The Okapi parameters were set as follows $k1$=1.4 b=0.8. For all our PRF runs, 5 documents were assumed relevant for term selection and document summaries comprised the best scoring 4 clusters. The *rsv* values to rank the potential expansion terms were estimated based on the top 20 or 40 ranked assumed relevant documents. The top 20 ranked expansion terms taken from the clusters were added to the original query in each case. Based on results from our previous experiments in CLEF, the original topic terms are up-weighted by a factor of 3.5 relative to terms introduced by PRF. For our submitted runs we used either the Title section (dcu*tit) or the Title and Description (dcu*desc) section of each topic. Our official submitted runs are marked $^+$ the tables of results. Baseline monolingual results using English topics without query expansion are given for comparison for each experimental condition.

For our experiments the document fields were combined as follows:
dcua2 – combination of ASRTEXT2004A and AUTOKEYWORDA1
dcua1a2 – combination of ASRTEXT2004A, AUTOKEYWORDA1 and AUTOKEYWORDA2
dcusum – combination of ASRTEXT2004A, AUTOKEYWORDA1 and AUTOKEYWORDA2 and the SUMMARY
dcuall – combination of ASRTEXT2004A, SUMMARY, NAME and MANUALKEYWORD

3.2 Experimental Results

Tables 1-4 show results of our experiments using these different data combinations for the 25 test topics released for the CLEF 2005 CL-SR task. Results shown are Mean Average Precision (MAP), total relevant documents retrieved (Rr), and

precision at cutoffs of 10 and 30 documents. Topic languages used are English, French, German and Spanish. Topics were translated into English using the Systran V3.0 machine translation system. The upper set of results in each table shows combined Title and Description topic queries and the lower set Title only topic queries.

Table 1. Results using a combination of ASRTEXT2004A and AUTOKEYWORDA1, with the Title or Title and Description topic fields. Expansion terms ranked for selection using statistics of 40 top ranked documents

Run-id	Topic Lang.	MAP	Rr	P10	P30
dcua2desc40f	Baseline	0.050	536	0.148	0.103
	English	0.065^{+}	738	0.176	0.140
	French	0.076	744	0.208	0.139
	German	0.041	611	0.116	0.099
	Spanish	0.055	727	0.152	0.109
dcua2tit40f	Baseline	0.070	384	0.228	0.143
	English	0.080	622	0.252	0.151
	French	0.081	708	0.252	0.155
	German	0.056	647	0.184	0.120
	Spanish	0.068	602	0.192	0.129

Results in Table 1 show results for combination of ASRTEXT2004A with AUTOKEYWORDA1. It can be seen that the PRF method improves results for the English topics in each case. Also that the results using Title only topics are better than those using the combined Title and Description topics with respect to MAP. This result is perhaps a little surprising since the latter are generally found to be perform better and we are investigating the reasons for the results observed here. However, the number of relevant documents retrieved is generally higher when using the combined topics which is to be expected since the topics will contain more terms which can match with potentially relevant documents. Cross-language information retrieval (CLIR) results using French topics are shown to perform better than monolingual English for both MAP and relevant retrieved. This is again unusual, but not unprecedented in CLIR. Results for translated German and Spanish topics show a reduction compared to the monolingual results.

Table 2. Results using a combination of ASRTEXT2004A, AUTOKEYWORDA1 and AUTOKEYWORDA2, with the Title or Title and Description topic fields. Expansion terms ranked for selection using statistics of 40 top ranked documents

Run-id	Topic Lang.	MAP	Rr	P10	P30
dcua1a2desc40f	Baseline	0.046	500	0.188	0.105
	English	0.067	784	0.184	0.148
	French	0.094	773	0.216	0.171
	German	0.046	611	0.096	0.092
	Spanish	0.064	765	0.164	0.128
dcua1a2tit40f	Baseline	0.0800	472	0.228	0.160
	English	0.110^{+}	727	0.252	0.196
	French	0.106^{+}	768	0.260	0.191
	German	0.074	691	0.172	0.149
	Spanish	0.091	679	0.220	0.156

Table 2 shows results for the same set of experiments as those in Table 1 with the addition of the AUTOKEYWORDA2 metadata to the documents. Results here generally show similar trends to those in Table 1 with small absolute increases in performance in most cases. In this case the performance advantage of French topics over English topics with PRF has largely disappeared for the Title only topics, however, performance for French topics is still much better than for English topics for the combined Title and Description topics.

Table 3. Results using a combination of ASRTEXT2004A, AUTOKEYWORDA1 and AUTOKEYWORDA2 and the SUMMARY section of each document, with the Title or Title and Description topic fields. Expansion terms ranked for selection using statistics of 40 top ranked documents

Run-id	Topic Lang.	MAP	Rr	P10	P30
dcusumdesc40f	Baseline	0.105	598	0.224	0.171
	English	0.147	889	0.272	0.217
	French	0.154	856	0.260	0.216
	German	0.108	696	0.164	0.137
	Spanish	0.107	860	0.168	0.152
dcusumtit40f	Baseline	0.141	618	0.284	0.216
	English	0.167	770	0.292	0.243
	French	0.165^{+}	837	0.308	0.251
	German	0.110	738	0.220	0.160
	Spanish	0.154	736	0.284	0.130

.

Table 3 shows results for a further set of experiments with the SUMMARY field added to the document descriptions. All results here show large increases compared to those in Table 2, indicating that the contents of the SUMMARY field are useful descriptions of the documents. The SUMMARY of each document is manually generated and presumably includes important terms which may be good descriptions of the topic of the document and possibly words actually appearing in the document, but incorrectly transcribed by the speech recognition system. The relative performance of monolingual and cross-language topics is the same as that observed in Table 2.

Table 4. Results using a combination of ASRTEXT2004A, SUMMARY, NAME and MANUALKEYWORD section of each document, with the Title or Title and Description topic fields. Expansion terms ranked for selection using statistics of 40 top ranked documents

Run-id	Topic Lang.	MAP	Rr	P10	P30
dcualldesc40f	Baseline	0.221	1031	0.368	0.271
	English	0.283	1257	0.432	0.337
	French	0.257	1122	0.424	0.303
	German	0.229	1001	0.328	0.272
	Spanish	0.247	1160	0.380	0.297
dcualltit40f	Baseline	0.242	736	0.412	0.311
	English	0.307	1009	0.488	0.377
	French	0.276	1136	0.496	0.360
	German	0.205	962	0.360	0.276
	Spanish	0.232	908	0.360	0.268

Table 4 shows a final set of experiments combining the ASRTEXT2004A, SUMMARY, NAME and MANUALKEYWORD fields. These results show large improvements over the results shown in previous tables. Performance for Title only and Title and Description combined topics is now similar with neither clearly showing an advantage. Monolingual English performance is now clearly better than results for translated French topics for both topic types, while our PRF method is still shown to be effective. The manually assigned keywords are shown to be particularly useful additional search fields.

4 Conclusions and Further Work

Our initial experiments with the CLEF 2005 CL-SR task illustrate that PRF can be successfully applied to this data set, and that the different fields of the document set make varying levels of positive contribution to information retrieval effectiveness. In general in can be seen that manual assigned fields are more useful than the automatically generated ones.

These experiments only represent a small subset of those that are possible with this dataset. In order to better understand the usefulness of document fields and retrieval methods more detailed analysis of these existing results and further experiments are

planned. The okapi retrieval model generally produces competitive retrieval results. However, in this case the results achieved are significantly lower than those observed using a parameter setting of the SMART retrieval system [7]. It is important to understand why the standard okapi weighting does not appear to work well with the CLEF 2005 CL-SR test collection, and we will be pursuing this issue as part of our further work.

References

1. White, R. W., Oard, D. W., Jones, G. J. F., Soergel, D., and Huang, X.: Overview of the CLEF-2005 Cross-Language Speech Retrieval Track, Proceedings of the CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation,Vienna, Austria, 2005.
2. Lam-Adesina, A. M., and Jones, G. J. F.: Applying Summarization Techniques for Term Selection in Relevance Feedback, Proceedings of the Twenty-Fourth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1-9, New Orleans, 2001. ACM.
3. Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M. ,and Gatford, M.: Okapi at TREC-3, Proceedings of the Third Text REtrieval Conference (TREC-3), pages 109-126. NIST, 1995.
4. Porter, M. F.: An Algorithm for Suffix Stripping, Program, 14:10-137, 1980.
5. Luhn. H.P.: The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2(2):159-165, 1958.
6. Jones, G. J. F., Burke, M., Judge, J., Khasin, A., Lam-Adesina, A. M., and Wagner, J.: Dublin City University at CLEF 2004: Experiments in Monolingual, Bilingual and Multilingual Retrieval, Proceedings of the CLEF 2004: Workshop on Cross-Language Information Retrieval and Evaluation, Bath, U.K., pages 207-220, 2004.
7. Tombros, A., and Sanderson, M.: The Advantages of Query-Biased Summaries in Information Retrieval. In proceedings of the Twenty-First Annual International ACM SIGIR Conference Research and Development in Information Retrieval, pages 2-10, Melbourne, 1998. ACM.
8. Inkpen, D., Alzghool, M., and Islam, A. : University of Ottawa's Contribution to CLEF 2005, the CL-SR Track Proceedings of the CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation, Vienna, Austria, 2005.